Estimation of Gait Relative Attribute Distributions using a Differentiable Trade-off Model of Optimal and Uniform Transports

Yasushi Makihara[1], Yuta Hayashi[1], Allam Shehata[1], Daigo Muramatsu[2,1], Yasushi Yagi[1]
[1]Osaka University [2]Seikei University
{makihara, hayashi, allam, yagi}@am.sanken.osaka-u.ac.jp muramatsu@st.seikei.ac.jp

Abstract

This paper describes a method for estimating gait relative attribute distributions. Existing datasets for gait relative attributes have only three-grade annotations, which cannot be represented in the form of distributions. Thus, we first create a dataset with seven-grade annotations for five gait relative attributes (i.e., beautiful, graceful, cheerful, imposing, and relaxed). Second, we design a deep neural network to handle gait relative attribute distributions. Although the ground-truth (i.e., annotation) is given in a relative (or pairwise) manner with some degree of uncertainty (i.e., inconsistency among multiple annotators), it is desirable for the system to output an absolute attribute distribution for each gait input. Therefore, we develop a model that converts a pair of absolute attribute distributions into a relative attribute distribution. More specifically, we formulate the conversion as a transportation process from one absolute attribute distribution to the other, then derive a differentiable model that determines the trade-off between optimal transport and uniform transport. Finally, we learn the network parameters by minimizing the dissimilarity between the estimated and ground-truth distributions through the Kullback–Leibler divergence and the expectation dissimilarity. Experimental results show that the proposed method successfully estimates both absolute and relative attribute distributions.

1. Introduction

Gait is one of the behavioral biometric modalities that is available at some distance from a camera and does not require subject cooperation. These properties have led to the use of gait in a variety of applications, such as surveillance and forensics using CCTV footage [1]. Indeed, gait recognition (i.e., gait-based identity recognition) has received attention from many researchers, and numerous gait recognition methods have been proposed over the past two decades [4].

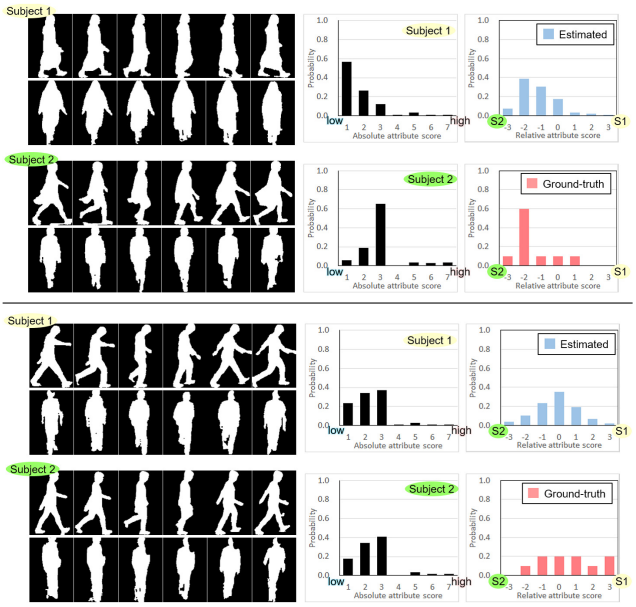


Figure 1: Examples of side-view and frontal-view gait silhouettes (left), estimated distributions of their corresponding absolute beautifulness scores (middle), and the estimated and ground-truth distributions of relative beautifulness scores for subjects 1 and 2 (right), where positive relative scores indicate that subject 1 scores higher than subject 2, and vice versa. The ground-truth distribution of the relative beautifulness is peaky for the upper pair (i.e., the annotators consistently score subject 2 higher), probably because of clear positive/negative cues (e.g., subject 2 has larger strides and more arm swing than subject 1). In contrast, the distribution has a greater spread for the lower pair (i.e., annotators' judgments are more inconsistent), probably because of mixed positive/negative cues (e.g., subject 1 has large arm swing, but appears to stoop). The proposed method can successfully estimate peaky/spread distributions for relative/absolute beautifulness.

Gait attributes include age [17, 15, 20], gender [27],

978-1-6654-3780-6/21/$31.00 ©2021 IEEE

emotions (e.g., happy and sad) [28], situations (e.g., nervous and relaxed) [24], soft biometrics (e.g., arm swing and stride) [19], and human perception-based attributes (e.g., beautiful, graceful, imposing) [21]. Of these, the human perception-based attributes are aesthetically important, because people who pay attention to their fashion style and body shape may also pay attention to their gait, i.e., whether their gait looks nice or not.

For example, consider the gait silhouette sequences of the upper pair in Fig. 1. We may perceive differences in beautifulness from their gait: subject 2 has a larger arm swing/stride than subject 1. However, it is not necessarily easy to check our gait by ourselves (e.g., by using an extremely large mirror, such as in a ballet classroom), whereas fashion style and body shape can be easily checked in the home using a standard-sized mirror. A video-based assessment tool of the aesthetic aspect of gait would therefore be useful for the self-training of one's gait (i.e., making one's gait more beautiful or graceful). Moreover, a witness to a criminal act may describe the gait of a suspect using human perception-based attributes (e.g., a person with an imposing gait walked away from the crime scene). In such cases, we could retrieve suspects by searching CCTV footage for people whose gait matches the witnessed attributes.

To realize such a gait attribute estimator, annotation data (i.e., gait video and its corresponding attribute labels) are essential, similar to general machine learning problems. However, we face some difficulties in annotating the gait attributes, although we can clearly annotate subject IDs for gait recognition and gender/ages for gait-based gender/age estimation. For example, assume that you are asked to annotate the beautifulness of gait of the subjects in Fig. 1. You may easily determine that subject 2 has a more beautiful gait than subject 1 in the upper pair, but may feel unable to assign each of them a specific score (e.g., 93 points out of 100).

The difficulties of attribute annotation are not confined to studies on gait, but also apply to faces and scenes. An alternative solution to annotation and its corresponding machine learning scheme are the so-called *relative attributes*, as proposed by Parikh and Grauman [18]. In the relative attribute framework, instead of annotating absolute attribute scores (or categorical labels) for each training sample, an annotator is shown a pair of training samples and assigns a relative attribute score for the pair (e.g., the first one is better, the two are similar, or the second one is better). For example, in the upper pair of Fig. 1, an annotator would be confident in annotating a relative score reflecting that "subject 2's gait is more beautiful than subject 1's gait." We then apply machine learning techniques to rank the annotated training samples (e.g., ranking support vector machine [12], deep ranking models [22, 26]), and derive an attribute estimator.

Although the annotation is given in a relative (or pairwise) manner in the relative attribute framework, note that the derived attribute estimator does not necessarily require a pair of inputs in the test phase. In other words, absolute attributes can be estimated from each input independently; otherwise, a user would need to find a partner to be compared with whenever he/she wished to use the attribute estimation system, which would be very laborious. This valuable property comes from the use of the same weights/parameters between pairs in the training phase (e.g., a Siamese network with shared parameters).

Taking a closer look at the annotation process, we notice that different annotators may assign different attribute labels to the same training sample because of the diversity of human perceptions on the attribute and/or the difficulty of annotating a particular training sample. To take this into consideration, we often ask multiple annotators to assign attribute labels (e.g., by a crowd sourcing service), and represent the ground-truth in the form of a distribution of the attribute rather than a single value, whereby the degree of uncertainty can be clearly seen. If we can estimate the uncertainty for a test sample, it could be used in several ways: a user would have some idea of the accuracy of the estimated attribute, or the system may reject the estimation result if the uncertainty is too high and ask the user to retry.

Uncertainty estimation itself has been studied for a long time, ranging from methods based on classical machine learning, e.g., Gaussian process regression [2], to those based on deep learning, e.g., label distribution learning [7], which are employed in applications such as facial age estimation [7, 10] and gait-based age estimation [20, 25]. These approaches are, however, only designed for absolute attributes, and hence cannot be directly applied to relative attributes. More specifically, a pair of inputs naturally outputs a pair of distributions of the absolute attribute, whereas the ground-truth is given in the form of a distribution of the relative attribute. Thus, we need to find a way to convert a pair of distributions of the absolute attribute into a single distribution of the relative attribute.

To overcome the abovementioned issue, this paper describes a method for estimating the relative attribute distribution. The contributions of this work are twofold.

1. Gait relative attribute datasets with finer-grade annotations. Existing datasets for gait relative attributes [21] contain only three-grade annotations, i.e., the first one is better, the two are similar, or the second one is better. Although these data are useful for relative attribute classification tasks, the three bins are insufficient for representing the distribution of gait attributes. Therefore, we create a dataset for gait relative attributes with standard seven-grade annotations assigned by multiple annotators.

2. A differentiable trade-off model of optimal and uniform transports is established to estimate the relative attribute distributions. As a key component in estimat-

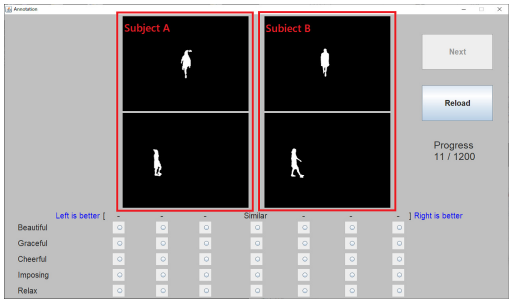


Figure 2: Screenshot of the seven-grade annotation tool for the five gait relative attributes.

ing the relative attribute distributions, we propose a model that converts paired distributions of absolute attributes to a single distribution of the relative attribute. Specifically, we regard the conversion as a transportation process from one absolute attribute distribution to the other, and then derive a differentiable model that determines the trade-off between the optimal transport, which assumes more coherent absolute attributes given by each annotator, and a uniform transport, which is formulated by entropy maximization.

2. Gait relative attribute dataset with seven-grade annotation

2.1. Annotation process

To decide the target gait attributes, we conducted preliminary annotation experiments. In these experiments, we asked several annotators to watch gait silhouette sequences from side and frontal views and provide gait attributes as freeform answers. We then consolidated these answers into the following five gait attributes: beautiful, graceful, cheerful, imposing, and relaxed.

We then chose walking videos of 1,200 subjects from OUMVLP [23]; to ensure the commonalty of attributes, all subjects were of a similar generation (i.e., aged in their twenties or thirties). Thereafter, we randomly chose 1,200 pairs from the 1,200 subjects under the constraint that each subject appears twice.

Next, we implemented a tool to annotate the gait attributes in a relative manner, as shown in Fig. 2. The annotation tool shows gait silhouette sequences of a pair of subjects from frontal and side views side-by-side. Each annotator was asked to select one of the seven grades for each attribute. The grades have values running from 3 (on the far left) to -3 (on the far right): grades 3, 2, and 1 indicate that the left sample is much better, better, or slightly better, respectively, grade 0 is neutral (i.e., the attributes are similar to each other), and grades -1, -2, and -3 indicate that the right sample is slightly better, better, or much better, respectively.

We employed ten annotators, and thus obtained ten annotations per pair. This gave a total of 12,000 annotations.

2.2. Statistics

We now present some statistics of the annotations. Specifically, we first discuss the average and standard deviation (SD) of the grades over 10 annotations for each sample and each attribute. Histograms of the average and SD over 1,200 pairs for each attribute are shown in Fig. 3. We can see that the averaged grades are distributed almost symmetrically around zero. The SD is mainly distributed between 0.5 and 2.0, which indicates the degree of uncertainty (or inconsistency among the annotators).

Moreover, the correlation of averaged grades between two attributes is shown in Fig. 4. For brevity, we only show the correlation between beautifulness and each of the other attributes. Beautifulness and gracefulness are highly correlated, while beautifulness and relaxedness are almost independent. The correlations between beautifulness and cheerfulness/imposingness lie somewhere in the middle.

3. Estimation of gait relative attribute distributions

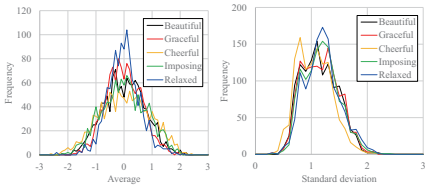
3.1. Overview

As discussed in the introduction, although annotation is given in a paired form, it is preferable for a gait attribute estimation system to output an absolute attribute from each subject's sample in a test case, rather than a relative attribute from a pair of subjects. Therefore, we design our framework so as to output a probability distribution of absolute attribute scores (referred to hereafter as an absolute distribution) from each subject's gait data, and then estimate a probability distribution of relative attribute scores (referred to hereafter as a relative distribution) from a pair of the absolute distributions. This is analogous to existing deep relative attribute approaches [22, 26, 9], which employ a Siamese network composed of two streams with parameter sharing.

Our framework is illustrated in Fig. 5. A side/front-view gait template is fed into a backbone network and its output feature vector is concatenated with auxiliary data. The result is then fed into a fully connected (FC) layer. The output feature vectors from side and frontal views are added and then normalized using the softmax operator to generate the absolute distribution. We set the number of nodes (scores or grades) of the absolute attributes to seven. Absolute distributions for an input pair (i.e., subjects 1 and 2 in this case) are further fed into the trade-off model to derive a relative distribution. Finally, the distribution dissimilarity and expectation dissimilarity are computed as a loss function.

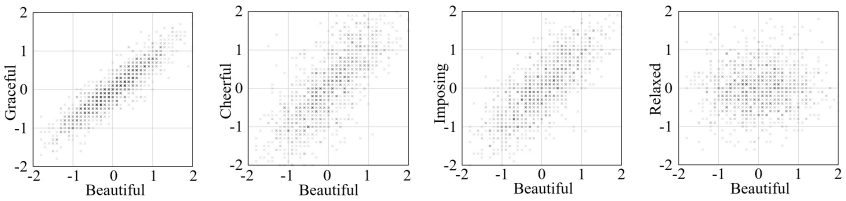
3.2. Input data

Because human perception-based gait attributes may not be sufficiently observed from a single view, we use gait data



(a) Average (b) SD

Figure 3: Histograms of average and standard deviation (SD) over annotators.



(a) Graceful (b) Cheerful (c) Imposing (d) Relaxed

Figure 4: Scatter plots of correlation between beautifulness and the other attributes.

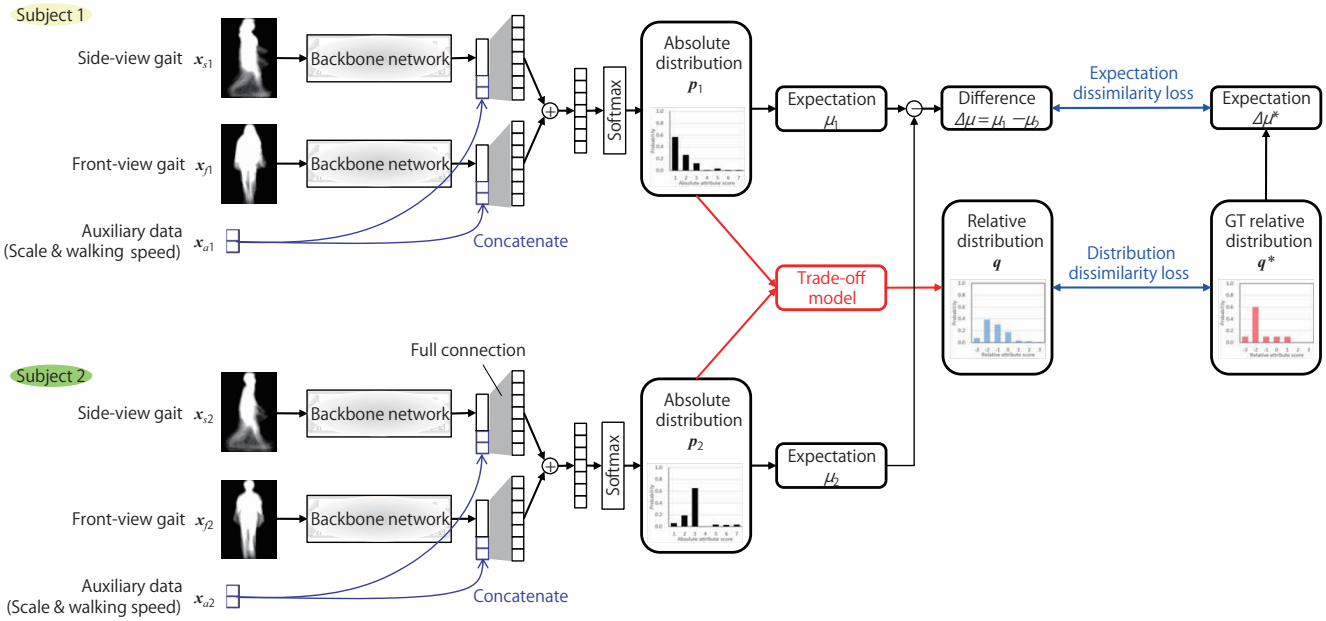


Figure 5: Overview of our framework. Given side/front-view gaits and auxiliary data, an absolute distribution is output via a backbone network and a fully connected (FC) layer. A pair of absolute distributions is then fed into the trade-off model to derive the relative distribution. Parameters for the backbone network and the FC layer are shared between views/subjects.

from two of the most representative views, i.e., side and frontal views. Specifically, we use side-view and frontal-view gait energy images [8], which are widely used gait representations. In addition, gait data are often provided in size-normalized and registered form (e.g., gait energy image), from which scale and translation information have been discarded. Therefore, we use the scale (i.e., the height before size normalization) and walking speed as auxiliary data; these data are obtained by computing the distance between the top of the head and the bottom of the feet. We also use the distance traveled during a certain time period in conjunction with ground-plane constraints and a calibrated camera in the world coordinates (e.g., [16]).

3.3. Backbone network

The backbone network has a conventional structure containing convolution, normalization, pooling, and FC layers, and its configuration is based on AlexNet [13], which is a standard network structure in the computer vision and pattern recognition community. Note that designing a novel network architecture is beyond the scope of this paper. Thus, other backbone networks such as GaitSet [3], GaitPart [6], and the gait lateral network [11] could also be used.

3.4. Trade-off model

3.4.1 Relative distribution by transport model

In this section, we describe how to estimate a relative attribute distribution from a pair of absolute attribute distributions output from each subject's stream. Recall that we only have the ground-truth for the relative attributes, rather than the absolute attributes. Thus, in the training phase, when measuring the dissimilarity between the ground-truth distribution and the estimated distribution for the loss function, it is essential to convert the pair of absolute attribute distributions to the relative attribute distribution.

As a preliminary step, we first consider a scalar attribute case. As demonstrated in previous work on relative attributes [18, 26, 22, 21], for a given pair of absolute attribute

scores (i.e., s_1 for subject 1 and s_2 for subject 2), we can easily compute the relative attribute in the form of the difference $\Delta s = s_1 - s_2$, and then employ some appropriate loss function such as the binary cross-entropy loss [22] or signed linear/quadratic contrastive loss [9].

However, estimating a relative attribute distribution given a pair of absolute attribute distributions is nontrivial. The main difficulty is as follows.

First, we introduce a set of absolute attribute scores in the integer domain as $\mathcal{S} = \{1, \ldots, S\}$, where S is the number of absolute attribute scores (or grades). We then introduce a discrete probability distribution of the absolute attribute scores for the i-th subject ($i = 1, 2$) of a pair as $\boldsymbol{p}_i = [p_i(1), \ldots, p_i(S)]^T$, where $p_i(s)$ is the probability of the s-th score for the i-th subject.

We then attempt to estimate a relative absolute distribution from the absolute attribute distributions $\boldsymbol{p}_1$, $\boldsymbol{p}_2$. Note that element-wise subtraction or division of the two distributions (e.g., $p_1(s) - p_2(s)$, $p_1(s)/p_2(s)$) does not make sense when estimating the relative attribute score/distribution. Therefore, we need to stick with score-level subtraction $s_1 - s_2$ to compute the relative attribute distribution, where s_1 and s_2 are drawn from $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$, respectively. To estimate the relative attribute distribution while maintaining the score-level subtraction principle, we consider a transportation process between the two absolute attribute distributions $\boldsymbol{p}_1$, $\boldsymbol{p}_2$. More specifically, we consider transporting the absolute attribute distribution for the first subject $\boldsymbol{p}_1$ to that for the second subject $\boldsymbol{p}_2$. Assuming that a portion of the probability $p_1(s_1)$ of score s_1 for the first subject is transported to the bin of score s_2 for the second subject (let the transportation amount be $m(s_1, s_2)(\geq 0)$), the difference between the absolute scores $\Delta s = s_1 - s_2$ and the transportation amount $m(s_1, s_2)$ are regarded as the relative attribute score and the portion of its corresponding probability that originated from the pair of absolute attribute scores (s_1, s_2). Based on this idea, we can compute the probability $q(\Delta s)$ of relative attribute score Δs by aggregating the possible pairs of absolute attribute scores as

$$q(\Delta s) = \sum_{(s_1, s_2) \in \mathcal{P}(\Delta s)} m(s_1, s_2), \tag{1}$$

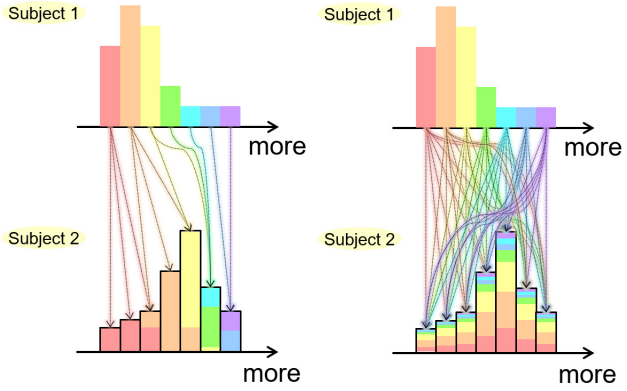
where $\mathcal{P}(\Delta s)$ is the set of pairs of absolute attribute scores of subjects 1 and 2 that have the difference Δs, defined as $\mathcal{P}(\Delta s) = \{(s_1, s_2) | s_1, s_2 \in \mathcal{S}, s_1 - s_2 = \Delta s\}$. Moreover, note that the transportation amounts are subject to the following two constraints so as to constitute the absolute attribute distribution for the first subject (referred to as the source distribution in the context of the transportation problem) $\boldsymbol{p}_1$ and the absolute attribute distribution for the second subject (referred to as the destination distribution) $\boldsymbol{p}_2$:

$$M\mathbf{1} = \boldsymbol{p}_1 \tag{2}$$

$$M^T\mathbf{1} = \boldsymbol{p}_2, \tag{3}$$

where $M \in \mathbb{R}^{N \times N}$ is an N by N matrix whose (s_1, s_2)-th component is $m(s_1, s_2)$ (referred to as the transportation matrix), and $\mathbf{1} \in \mathbb{R}^N$ is an N-dimensional one-padded vector.

In summary, the remaining problem for estimating the relative attribute distribution is how to estimate the transportation matrix M under the constraints of Eqs. (2) and (3) given the two absolute attribute distributions $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$. Because the transportation process is not unique (i.e., M is indefinite), we introduce two typical models (or assumptions) to constrain them in the following subsections.



(a) Optimal transport model (b) Uniform transport model

Figure 6: Concept of transport models. The optimal transport model (a) makes the transport process more coherent by minimizing the transportation cost, while the uniform transport model (b) makes the transport process more diverse by maximizing the entropy of the transport.

3.4.2 Optimal transport model

The first model assumes that the virtual annotation labels are coherent on absolute attribute scores that are never explicitly obtained, i.e., the scores come to each annotator's mind when annotating a relative attribute label. In other words, we assume that one annotator may tend to consistently assign higher absolute attribute scores, while another tends to assign lower absolute attribute scores. For example, if annotator A assigns a higher absolute attribute score than annotator B for the first subject of an input pair, the model expects that annotator A will also assign a higher absolute attribute score than annotator B for the second subject of the pair, and vice versa.

Therefore, this model tends to transport the probability of a certain bin in the source distribution to as similar a bin as possible in the destination distribution (see Fig. 6

(a)). This is known as the optimal transport process, and is formulated as the following transportation minimization problem under the constraints of Eqs. (2) and (3):

$$C(\boldsymbol{M}) = \sum_{s_1=1}^{S} \sum_{s_2=1}^{S} c(s_1, s_2) m(s_1, s_2) \rightarrow \min, \quad (4)$$

where $c(s_1, s_2)$ is a transportation cost function for scores s_1 and s_2, defined as the squared difference of the scores, $c(s_1, s_2) = (s_1 - s_2)^2$.

The abovementioned transportation minimization problem can be solved using the Hungarian algorithm [14]. This, however, involves linear programming, and is therefore not a differentiable process, which is critical in the backpropagation stage associated with the training of a deep neural network.

3.4.3 Uniform transport model

The second model assumes the independence of the virtual annotation labels on absolute attribute score. More specifically, each annotator assigns an absolute attribute score for the second subject of an input pair independently of the absolute attribute score assigned to the first subject of the pair. In other words, we assume that the annotators assign absolute attribute scores independently for each sample in a pair.

This model tends to transport the probability of a certain bin in the source distribution to all bins of the destination distribution as uniformly as possible (see Fig. 6 (b)). The uniform transport process is formulated as an entropy maximization process under the constraints of Eqs. (2) and (3) as

$$H(M) = -\sum_{s_1=1}^{S} \sum_{s_2=1}^{S} m(s_1, s_2) \ln m(s_1, s_2) \rightarrow \max. \quad (5)$$

This is equivalent to the case in which each sample is independently drawn from the source and destination distributions, and so the transportation amount (i.e., a kind of joint probability of scores s_1 and s_2) is simply represented by $m(s_1, s_2) = p_1(s_1)p_2(s_2)$. This is naturally differentiable in terms of the probability of the source and destination distributions, and is therefore suitable for a deep learning framework.

3.4.4 Trade-off model

Because the abovementioned models favor the two extreme properties of coherence and independence of the scores, we use a trade-off model. The trade-off model is referred to as entropy-regularized optimal transport [5], and is formulated by the following weighted sum of Eqs. (4) and (5):

$$L(M) = C(M) - \varepsilon H(M) \rightarrow \min, \quad (6)$$

where ε is a hyperparameter that controls the entropy regularization. The entropy-regularized optimal transport can be solved by the efficient Sinkhorn–Knopp algorithm [5]. More specifically, by solving Eq. (6) under the constraints of Eqs. (2) and (3) with Lagrange multipliers, we obtain the following alternate solutions:

$$\boldsymbol{u}_2 = \boldsymbol{p}_2 ./ (K^T \boldsymbol{u}_1) \quad (7)$$

$$\boldsymbol{u}_1 = \boldsymbol{p}_1 ./ (K \boldsymbol{u}_2) \quad (8)$$

$$M = \mathrm{diag}(\boldsymbol{u}_1) K \mathrm{diag}(\boldsymbol{u}_2), \quad (9)$$

where $K \in \mathbb{R}^{S \times S}$ is an S by S matrix whose (s_1, s_2)-th component is $\exp(-c(s_1, s_2)/\varepsilon)$, $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^S$ are S-dimensional temporary vector variables, $./$ indicates an element-wise division operator, and $\mathrm{diag}(\cdot)$ indicates a function that converts a vector into a diagonal matrix. After initializing $\boldsymbol{u}_1$ to a one-padded vector, i.e., $\boldsymbol{u}_1 = \boldsymbol{1}$, we alternately update $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ using Eqs. (7) and (8) several times, and finally compute the transportation matrix M from Eq. (9). As seen from Eqs. (7), (8), and (9), the Sinkhorn–Knopp algorithm simply involves iterative linear algebraic computations, and hence guarantees that differentiability is retained for the backpropagation process. This enables us to employ the trade-off model in a deep learning framework, unlike the optimal transport model alone.

3.5. Loss function

We introduce a final loss function composed of an expectation dissimilarity and distribution dissimilarity. As shown in Fig. 5, once the network outputs the paired distributions of absolute attribute scores $\boldsymbol{p}_1$, $\boldsymbol{p}_2$, we compute their means μ_1, μ_2, take the difference $\Delta\mu = \mu_1 - \mu_2$, and then form a squared error using the ground-truth expectation $\Delta\mu^*$ as $L_{\mathrm{EX}} = (\Delta\mu - \Delta\mu^*)^2$. We further compute the Kullback–Leibler (KL) divergence between the relative attribute score distribution $\boldsymbol{q}$ estimated by our trade-off model and the ground-truth distribution $\boldsymbol{q}^*$ as

$$L_{\mathrm{KL}}(\boldsymbol{q}^* || \boldsymbol{q}) = \sum_{\Delta s} q^*(\Delta s) \log \left(\frac{q^*(\Delta s)}{q(\Delta s)} \right). \quad (10)$$

This gives the distribution dissimilarity.

Finally, we minimize their weighted sum as

$$L = w_{\mathrm{KL}} L_{\mathrm{KL}} + w_{\mathrm{EX}} L_{\mathrm{EX}}, \quad (11)$$

where w_{KL} and w_{EX} are hyperparameters that control the KL divergence and the expectation error.

4. Experiments

4.1. Setup

We used the dataset described in section 2 to conduct experimental evaluations. We extracted 900 training pairs

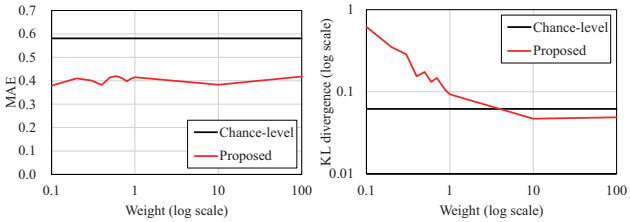
from 1,000 subjects and 100 test pairs from the other 200 subjects. Note that each subject appears twice in the 1,200 pairs in the dataset, and so the remaining 200 pairs were discarded so that the subjects in the training and test sets remained completely disjoint. We trained the network parameters for each gait attribute separately over 100 epochs using a learning rate of 0.0001. We experimentally set the entropy regularization weight $\varepsilon = 10$ and the expectation and KL divergence weights as $w_{\mathrm{EX}} = 0.7$ and $w_{\mathrm{KL}} = 0.3$, respectively. We considered seven grades for the absolute scores (i.e., $S = 7$) and clipped the relative scores to be within the seven-grade relative score scale from -3 to $+3$, i.e., we reformulated the relative scores as $\Delta s = \min(\max(s_1 - s_2, -3), 3)$ in practice. We evaluated the accuracy of the expectation and distribution through the mean absolute error (MAE) and KL divergence, respectively.

4.2. Qualitative evaluation

Samples with high and low estimated absolute beautifulness scores are presented in Fig. 1 for the purpose of qualitative evaluation. The ground-truth distribution of the relative beautifulness is peaky for the upper pair, possibly because of clear positive/negative clues (e.g., subject 2 has larger strides and greater arm swing than subject 1), while the distribution of the lower pair has a greater spread, possibly because of mixed positive/negative clues (e.g., subject 1 has large arm swing, but a greater tendency to stoop, while subject 2 has mid-range arm swing and stoop). In both cases, the proposed method successfully estimates peaky/spread distributions for relative/absolute beautifulness. For example, the absolute distributions for subjects 1 and 2 for the upper pair have clear peaks around $s = 1$ and $s = 3$, respectively, whereas those for the lower pair do not have clear peaks, which indicates that the system accurately perceives the uncertainty (or confidence) in the estimation of the absolute attributes.

4.3. Quantitative evaluation

The accuracy of the estimated expectation and distributions was quantitatively evaluated through comparisons with a baseline method of score regression, which minimizes the mean squared error (MSE) between the estimated score and the ground-truth. Because the MSE does not estimate a distribution and there is no way of estimating a relative attribute distribution, we include the chance-level performance as another baseline—this returns a prior relative attribute distribution obtained from the training data. We report the MAEs and KL divergence as expectation and distribution accuracy measures in Table 1. The results show that the proposed method is the most accurate in terms of both MAE and KL divergence in all but one cases.



(a) MAE (b) KL divergence

Figure 7: Sensitivity analysis of MAE and KL divergence on the entropy weight for beautifulness.

4.4. Sensitivity analysis

Finally, we conducted a sensitivity analysis of the entropy regularization weight ε for both the MAE and KL divergence. The results are shown in Fig. 7. The MAE is relatively insensitive to ε, whereas the KL divergence is sensitive to relatively small values (e.g., $\varepsilon < 10$). This is because ε tends to make the estimated relative distribution less uniform, and so the probabilities for some of the bins are quite small (i.e., $q(\Delta s) \approx 0$). Consequently, the KL divergence increases significantly.

5. Conclusion

This paper has addressed the estimation of gait relative attribute distributions. We first constructed a gait relative attributes dataset with seven-grade annotations, which is suitable for representing distributions. We then proposed a trade-off model that converts paired absolute distributions into a relative distribution based on optimal and uniform transport processes.

One avenue of future research is to extend the proposed method to a more unified framework in which an entropy regularization parameter can be learned in an end-to-end manner. Another direction involves investigating other loss functions (e.g., Jensen–Shannon divergence) to provide a better trade-off between the optimal and uniform transport processes. Additionally, we will apply the proposed method to other biometric modalities such as faces for further validation.

Acknowledgment This work was supported by JSPS KAKENHI Grant No. JP18H04115, JP19H05692, and JP20H00607. We thank Stuart Jenkinson, PhD, from Edanz Group for editing a draft of this manuscript

References

[1] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon. On using gait in forensic biometrics. *Journal of Forensic Sciences*, 56(4):882–889, 2011.

[2] C. K. I. W. Carl Edward Rasmussen. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[3] H. Chao, Y. He, J. Zhang, and J. Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings*

Table 1: MAE and KL divergence. Bold font indicates the best accuracy.

Measure	Method	Beautiful	Graceful	Cheerful	Imposing	Relaxed	Average
MAE	Chance-level	0.581	0.493	0.769	0.643	0.498	0.597
	MSE	0.388	0.364	0.405	0.417	**0.315**	0.378
	Proposed	**0.383**	**0.350**	**0.369**	**0.398**	0.322	**0.364**
KL divergence ($\times 10^{-2}$)	Chance-level	6.178	5.146	7.817	6.262	5.616	6.204
	Proposed	**4.729**	**4.212**	**4.991**	**4.768**	**4.483**	**4.637**

of the AAAI Conference on Artificial Intelligence, volume 33, pages 8126–8133, Jul. 2019.

[4] P. Connor and A. Ross. Biometric recognition by gait: A survey of modalities and features. *Computer Vision and Image Understanding*, 167:1–27, 2018.

[5] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[6] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He. Gaitpart: Temporal part-based model for gait recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14213–14221, Jun. 2020.

[7] B. Gao, C. Xing, C. Xie, J. Wu, and X. Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017.

[8] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(2):316– 322, Feb. 2006.

[9] Y. Hayashi, A. Shehata, Y. Makihara, D. Muramatsu, and Y. Yagi. Deep gait relative attributeusing a signed quadratic contrastive loss. In *Proc. of the 25th Int. Conf. on Pattern Recognition (ICPR 2020)*, pages 1–8, Jan. 2021.

[10] Z. He, X. Li, Z. Zhang, F. Wu, X. Geng, Y. Zhang, M. Yang, and Y. Zhuang. Data-dependent label distribution learning for age estimation. *IEEE Transactions on Image Processing*, 26(8):3846–3858, 2017.

[11] S. Hou, C. Cao, X. Liu, and Y. Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *Prof. of the 16th European Conf. on Computer Vision (ECCV 2020)*, pages 382–398, Cham, Aug. 2020. Springer International Publishing.

[12] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, 2002.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*.

[14] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1 and 2):83–97, 1955.

[15] J. Lu and Y.-P. Tan. Ordinary preserving manifold analysis for human age and head pose estimation. *IEEE Transactions o Human-Machine Systems*, 43(2):249–258, March 2013.

[16] Y. Makihara, G. Ogi, and Y. Yagi. Geometrically consistent pedestrian trajectory extraction for gait recognition. In *Proc. of the IEEE 9th Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS 2018)*, pages 1–11, Oct. 2018.

[17] Y. Makihara, M. Okumura, H. Iwama, and Y. Yagi. Gait-based age estimation using a whole-generation gait database. In *Proc. of the Int. Joint Conf. on Biometrics (IJCB2011)*, pages 1–6, Washington D.C., USA, Oct. 2011.

[18] D. Parikh and K. Grauman. Relative attributes. In *2011 International Conference on Computer Vision*, pages 503–510. IEEE, 2011.

[19] D. A. Reid and M. S. Nixon. Using comparative human descriptions for soft biometrics. In *Proc. of the 1st Int. Joint Conf. on Biometrics (IJCB 2011)*, pages 1–6, 2011.

[20] A. Sakata, Y. Makihara, N. Takemura, D. Muramatsu, and Y. Yagi. How confident are you in your estimate of a human age? uncertainty-aware gait-based age estimation by label distribution learning. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2020.

[21] A. Shehata, Y. Hayashi, Y. Makihara, D. Muramatsu, and Y. Yagi. Does my gait look nice? human perception-based gait relative attributes estimation by dense trajectory analysis. In *Proc. of the 5th Asian Conf. on Pattern Recognition (ACPR 2019)*, pages 1–14, Nov. 2019.

[22] Y. Souri, E. Noury, and E. Adeli. Deep relative attributes. In *Asian conference on computer vision*, pages 118–133. Springer, 2016.

[23] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1):4, Feb 2018.

[24] N. F. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 2:371–387, Sep. 2002.

[25] C. Xu, Y. Makihara, R. Liao, H. Niitsuma, X. Li, Y. Yagi, and J. Lu. Real-time gait-based age estimation and gender classification from a single image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3460–3470, January 2021.

[26] X. Yang, T. Zhang, C. Xu, S. Yan, M. S. Hossain, and A. Ghoneim. Deep relative attributes. *IEEE Transactions on Multimedia*, 18(9):1832–1842, 2016.

[27] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu. A study on gait-based gender classification. *IEEE Trans. on Image Processing*, 18(8):1905–1910, Aug. 2009.

[28] Y. Zhuang, L. Lin, R. Tong, J. Liu, Y. Iwamot, and Y.-W. Chen. G-gcsn: Global graph convolution shrinkage network for emotion perception from gait. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.